*International Journal of Electrical and Computing Engineering*

*Vol. 1, Issue. 2, December – 2014* *ISSN (Online): 2349-8218*

# Band Selection Using Clustering Technique for Dimensionality Reduction in Hyper spectral Image

**[1]Karthick.V, [2]Veera Senthil Kumar.G, [3]Dr. Vasuki. S**

[1]Lecturer, [2]Assistant Professor, [3]Professor and Head, [1,2,3] Dept.of.ECE,
Velammal College of Engineering and Technology, Madurai, INDIA,
vkt@vcet.ac.in, gvs@vcet.ac.in, sv@vcet.ac.in

***Abstract* - This paper presents a novel approach for band selection using K-means clustering technique for dimensionality reduction in Hyperspectral images (HSI). Despite many algorithms exist for dimensionality reduction, it is even now a challenging task of selecting informative bands from the large volume data. The number of bands is estimated with the concept of Virtual Dimensionality (VD), because it provides reliable estimate. Bands are clustered using K-means based on the statistical measures such as Variance(VAR), Standard Deviation(STD) and Mean Absolute Deviation(MAD). Bands preserving maximum information are selected based on the maximum value of statistical measures. Finally, End members are extracted from the selected bands using N-FINDR and their spectral signatures are determined. The whole experimentation is carried out in MATLAB.**

***Keywords- Dimensionality Reduction, K-means clustering, VD, N-FINDR.***

## I. INTRODUCTION

Hyperspectral image is a collection of number of spectral bands where each image pixel is represented by a column vector where each of column components is a pixel imaged by a particular spectral channel. The collected image data by hyperspectral remote sensors is simultaneously in hundreds of narrow, adjacent spectral bands over the wavelengths that can range from the near ultraviolet through the thermal infrared at 5nm of fine resolutions. Each pixel contains a hyperspectral signature that represents different materials. As a result of high spectral resolution, hyperspectral systems produce a massive amount of data. These measurements make it possible to derive a continuous spectrum for an image data. Hyperspectral data helps the analyst in detection of more materials, objects and regions with enhanced accuracy.

Hyperspectral images provide a vast amount of information about a scene, but most of that information is redundant as the bands are highly correlated. For computational and data compression reasons, it is desired to reduce the dimensional of the data set while maintaining good performance in image analysis tasks. There are some of the challenges to be faced during the analysis of hyperspectral images. First issue is data storage and transmission problem due to huge data volume. Second one is redundancy of information because redundancy in data can cause convergence instability. Third one is concerned with high processing time. As a result, the imposition of requirements for storage space, computational load and communication bandwidth are against the real time applications and it is difficult to visualize or to classify such a huge amount of data. Dimensionality reduction is a good choice to overcome these challenges. The reduction of dimensionality is necessary for high accuracy in unmixing of the pixels, classification and detection.

There are several methods of dimensionality reduction which can be further categorized into two groups; feature extraction and feature or band selection. Feature selection is preferable for dimensionality reduction because feature extraction need most of the original data representation for extraction of features. Secondly due to transformation in feature extraction the critical information may have been distorted. Compared to feature extraction, feature selection preserves the relevant original information.

The overview of this paper is as follows: the details of the AVIRIS [1] hyperspectral image JASPER RIDGE is discussed in the section II. Estimation of number of bands by the concept of VD [1] is described in the section III. Band selection using K-means Clustering algorithm [3] is explained in section IV. N-FINDR [2], an end member extraction algorithm to detect end members is discussed in section V. The

proposed methodology is explained in section VI. Experimental results are discussed in section VII. Conclusion is given in section VIII.

## II. HSI DATA

Jasper Ridge is a AVIRIS [1] hyperspectral data used for our experimentaion. There are 512 x 614 pixels in it. Each pixel is recorded at 224 channels ranging from 380 nm to 2500 nm. The spectral resolution is up to 9.46nm. Since this hyperspectral image is too complex to get the ground truth, subimage is considered of 100 x 100 pixels. The first pixel starts from the (105,269) pixel in the original image.

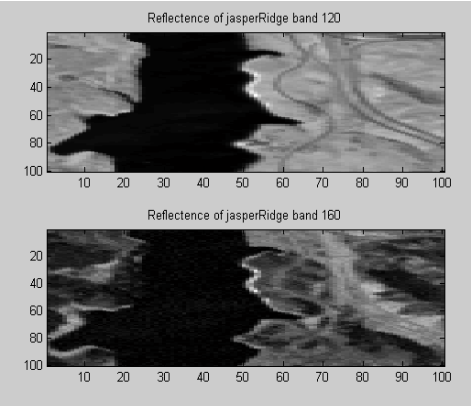



Fig. 1 Reflectance of Jasper Ridge image band no. 120 &160

After removing the channels 1--3, 108--112, 154--166 and 220--224 due to dense water vapor and atmospheric effects, 198 channels remain for further analysis. Out of 198 channels , reflectances of Jasper Ridge image band 120,160 are shown in Fig. 1. Jasper ridge is an image with 10000 samples and 195 dimensions are selected. 100% of eigen values are retained. Road ,Tree, Soil and Water are the four end members, found in the ground truth image. Their spectral signatures are shown in Fig. 2.

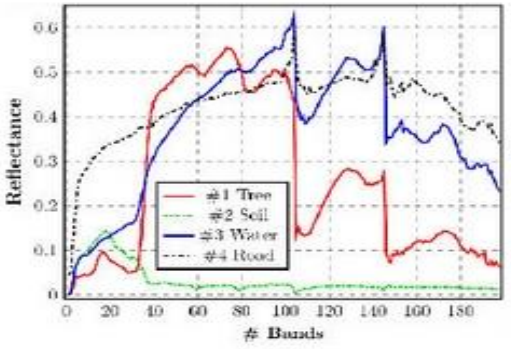



Fig. 2 Ground truth of endmembers spectral signatures

## III. VIRTUAL DIMENSIONALITY

Virtual Dimensionality (VD) [1] estimates the number of endmembers present in the HSI image. Two familiar methods widely used for estimating VD are Harsanyi-Farrand-Chang (HFC) method as well as a noise whitened version (NWHFC)] [5], on the basis of Neyman-Pearson detection theory to explore how many times the test fails for all spectral bands and for a given false-alarm probability, $P_F$. Since the HFC method does not have a noise-whitening process, an alternative is to modify the HFC method by including a noise-whitening process as preprocessing to remove the second-order statistical correlation such that the noise variance in the corresponding correlation eigen value and covariance eigen value will be the same. As a result, the VD estimate can be more accurate due to the fact that the noise variances have been decorrelated and do not have effects on the eigen value comparison. The resulting HFC method will be referred to as noise-whitened HFC method.

## IV. BAND SELECTION

Band Selection is performed using clustering technique. K-means clustering is adopted here because it is a familiar and simple unsupervised clustering technique. Clustering of band images are performed by keeping the intra-cluster variance minimum and the inter-cluster variance maximum. The method in which dimensionality is reduced by selecting a subset of the original dimensions are known as band/ feature selection. The hyperspectral data is spread in some direction. This data can be measured by using different statistical methods which include MAD [1], moment, variance, mean, geometric mean and standard deviation. MAD, STD and VAR [1] are used here for measuring the data. Suppose that we have $\{B_l\}_{l=1}^{L}$ band images in our hyperspectral image data cube where $\mathbf{L}$ is the total number on bands, if each band image is of size $\mathbf{M \times N}$ and $\bar{B}_l$ the mean of the band image. The statistical characteristics used for data are as follows.

MAD for the $l^{\text{th}}$ band is $d_l = \frac{1}{MN}\sum_{i=1}^{MN}\left|b_i - \bar{B}_l\right|$

Standard Deviation for the band image is

$$d_l = \left(\frac{1}{MN}\sum_{i=1}^{MN}\left(b_i - \bar{B}_l\right)^2\right)^{\frac{1}{2}}$$

Variance for the band image is

$$d_l = \frac{1}{MN}\sum_{i=1}^{MN}\left(b_i - \overline{B}_l\right)^2$$

The result from the above statistical methods for L band images is given by:

$$d = \{d_l\}_{l=1}^{L}$$

K-means clustering is one of the simplest unsupervised algorithms and is well-known for solving the problem of clustering. The flowchart of K-means clustering is shown in Fig. 3. K-means follows a simple and easy way to classify a given data set through clusters; the number of clusters is fixed and is given a prior. The number of centroids i.e. K are defined for each cluster and which are placed far away from each other as possible. The points which belong to the given data set are taken and are associated to the nearest centroid which results in K number of groups. Again K new centroids are recalculated for new centers of the cluster and a new binding has to be done between the same data set points and the nearest new centroid. A loop is run for the K centroids to change their location step by step until there is no change and the centroids are fixed. The centroids of the clusters are calculated by minimizing the sum of squared errors. The K means algorithm performs three steps until convergence.

1) Determine the centroid coordinate
2) Determine the distance of each object to the centroids
3) Group the object based on minimum distance

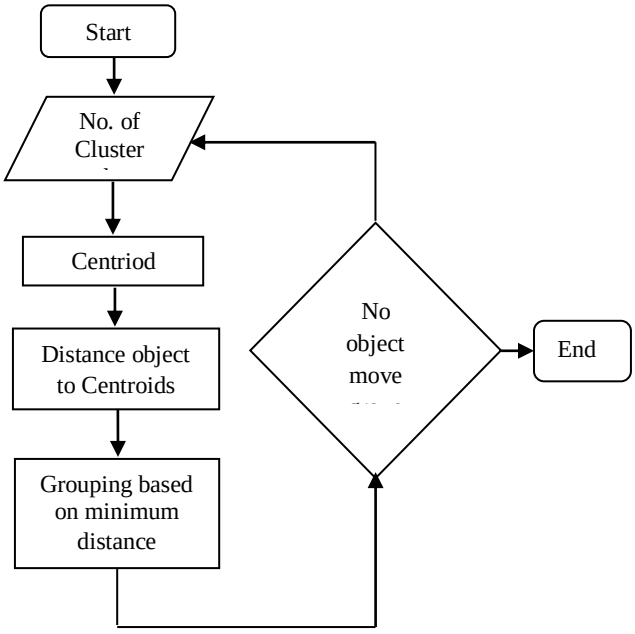



Fig. 3. K-Means clustering flow chart

(1). For the observation $X = (x_1, x_2, x_3....x_n)$, the K-means clustering method divides the n observations into k sets (k<n),k= $\{s_1, s_2, s_3,...s_k\}$, minimizing the sum of squares with-in clusters i.e.

(2). $\min_s \sum_{i=1}^{k}\sum_{x_j \in s_i}\left\|x_j - \mu_i\right\|$

Where $\mu_i$ is the mean of points in clusters $C_k$

(3). K-means computes centroid clusters using Squared Euclidean distance metric.

For an m-by-n data matrix $X = (x_1, x_2, x_3,...x_m)$ the distance between the vector $x_r$ and $x_s$ is given by

$$d_{rs}^2 = (x_r - x_s)D^{-1}(x_r - x_s)$$

where D is the diagonal matrix.

Bands are clustered based on their statistical characteristics i.e. Variance, MAD (Mean Absolute Deviation) and Standard Deviation by K-means clustering technique. A band is selected from each cluster which has maximum variance with in the cluster. The proposed technique using Variance with Squared Euclidean as distance metric is abbreviated as VAR-SE. Similarly for Standard Deviation with Squared Euclidean as STD-SE and the technique using MAD with Squared Euclidean is abbreviated as MAD-SE

## V. ENDMEMBER EXTRACTION

An endmember can be defined as an idealized, pure signature for a class. In hyperspectral data analysis, a pure pixel refers to an L dimensional pixel vector. Also it should be noted that an endmember is not a pixel. It is a spectral signature that is completely specified by the spectrum of a single material substance. There are two ways by which endmembers can be identified: Endmember Extraction Algorithm (EEA), which extract pure pixels directly from the data and Endmember Generation Algorithm(EGA), which aimed at generating pure signatures from available pure pixels. Here EEA is focused.

One of the most widely used EEAs has been the N-FINDR algorithm developed by Winter. It is an iterative simplex volume expansion approach which assumes that, in L spectral dimensions, the L dimensional volume formed by a simplex with vertices specified by purest pixels is always larger than that formed by any other combination of pixels. The generic implementation form of N-FINDR finds those vertices by randomly selecting a set of p pixels from the scene as initial endmembers, and calculating the volume of

the simplex formed by these initial endmembers. This process is iterated through the following steps to test every pixel in the image as an endmember. First, each of the initial endmembers is replaced one at a time with the pixel being tested. Second, the volumes of the simplexes formed by each replacement are calculated. Finally, the algorithm evaluates if replacing any of the initial endmembers with the pixel being tested results in a larger simplex volume. If this is the case, the pixel being tested replaces the initial endmember and the process is repeated again until each pixel is evaluated as a potential endmember. The pixels which remain as endmembers at the end of the process are considered to be the final endmembers.

## VI. Proposed Methodology

Hyperspectral imagecube with high dimensionality is preprocessed to remove all bands that are affected by dense water vapor and atmospheric effects. This is a common preprocess required in hyperspectral data analysis. After preprocessing, bands with high SNR is retained as good bands for further processing. The next step is dimensionality reduction inorder to reduce the computational complexity. Band selection strategy is adopted in this paper for dimensionality reduction. VD is estimated using NWHFC method with false alarm rate of $10^{-4}$, to calculate the number of bands required for endmember extraction. Band selection is done using K-means clustering based on the statistical measures of the input hyperspectral bands such as VAR, STD and MAD. End members are extracted using N-FINDR algorithm and their spectral signatures are determined.

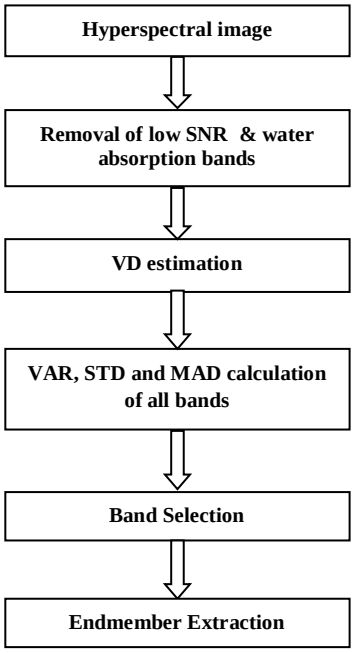



Fig. 4. Flow Diagram of Proposed Methodology

The steps of the proposed algorithm are summarized as follows:

1) Low SNR and water absorption bands from Hyperspectral imagecube is removed.
2) VD is estimated to know the number of bands required using NWHFC method.
3) The data of each band image is calculated using VAR, MAD and STD.
4) Bands are clustered using K-means clustering based on the measured values to examine the proximity of band images to each other.
5) According to VD, clusters are created which contain all the measured values.
6) Band having maximum value from each cluster is selected.
7) The indices of endmembers are determined using N-FINDR algorithm and the corresponding endmembers' signatures are plotted.

The complete process involved in the proposed methodology is depicted in Fig. 4.

## VII. Experimental Results

AVIRIS hyperspectral data JASPER RIDGE of size 100x100 pixels having 224 bands, is used for our experimentation. After removing low SNR and water absorption bands, 198 bands are preserved. The ground truth of endmembers spectral signatures shown in Fig.1 reveals that there exists four endmembers ie) Road, Tree, Soil and Water, in the image scene.

Preserving the maximum information, the number of bands required and estimated VD are 10 with false alarm rate of $10^{-4}$ using NWHFC method. The estimated VD for different false alarm rate($P_{F)}$ is shown in Table 1.

TABLE 1. ESTIMATE OF VD USING NWHFC METHOD FOR DIFFERENT FALSE ALARM RATE

| $P_F$ | $10^{-1}$ | $10^{-2}$ | $10^{-3}$ | $10^{-4}$ | $10^{-5}$ |
|---|---|---|---|---|---|
| **VD** | 21 | 17 | 12 | 10 | 9 |

Because of minor variations found in the spectral signatures, the number of bands is restricted to 4. Variance, Standard Deviation and Mean Absolute Deviation are calculated for 198 bands. Using these measures, bands are clustered using K-means algorithm. The number of clusters is based on the number of endmembers to be identified. Bands having maximum value are selected from each cluster and thus the band selection is achieved here. The selected bands based on different measures are shown in Table 2.

TABLE 2. SELECTED BANDS FOR DIFFERENT MEASURES

| Criteria | Selected Bands |
|---|---|
| VAR-SE | 182,118, 53,104 |
| STD-SE | 34,178, 41,104 |
| MAD-SE | 34,115, 51,100 |

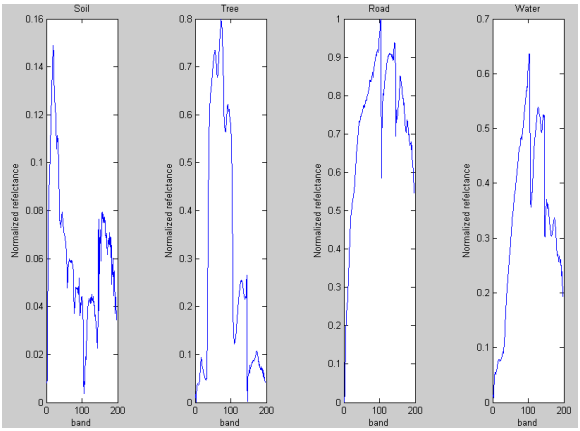



Fig. 5. Spectral Signatures of 4 endmembers using VAR

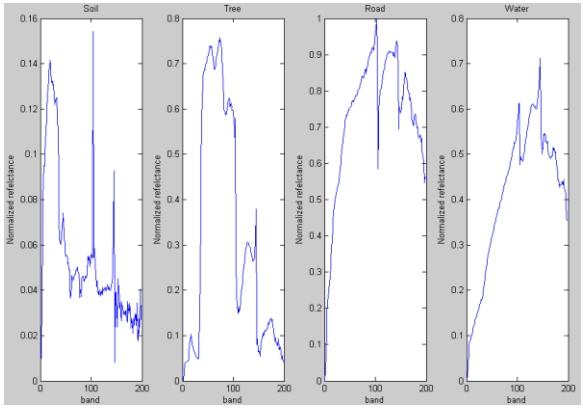
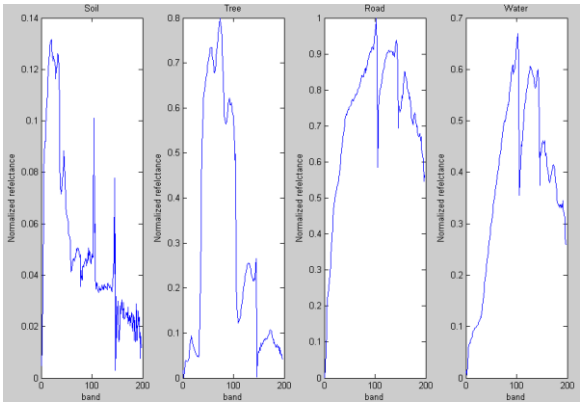



Fig. 6. Spectral Signatures of 4 endmembers using STD




The endmembers are extracted using N-FINDR algorithm and their spectral signatures are plotted in Fig. 5, Fig. 6, and Fig. 7.

Fig. 7. Spectral Signatures of 4 endmembers using MAD

## VIII. Conclusion

The proposed method of dimensionality reduction using K-means clustering provides better band selection. Further, the bands are selected using K-means clustering based on the various measures such as VAR, STD and MAD. In proposed technique of band clustering and selection using K-means method, band from each cluster is selected such that intra-cluster variance is kept maximum and inter-cluster variance is minimum. The proposed technique is simple to implement and computes the result very fast. The computation takes seconds for band clustering and selection. The experimental results show that the endmembers are detected well and the spectral signatures of endmembers found using N-FINDR algorithm are highly matched with the spectral signatures of endmembers shown in ground truth. Therefore it is concluded from the results of experiments that the proposed clustering techniques are promising and authentic techniques for band clustering and band selection.

## IX. References

[1] Muhammad Sohaib, Ihsan-Ul-Haq, and Qaiser Mushtaq, "Dimensional Reduction of Hyperspectral Image Data Using Band Clustering and Selection Based on Statistical Characteristics of Band Images", *International Journal of Computer and Communication Engineering, Vol. 2, No. 2, March 2013*

[2] Antonio Plaza and Chein-I Chang, "An Improved N-FINDR Algorithm in Implementation", Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XI, Proceedings of SPIE Vol. 5806

[3] Chein-I Chang, *Senior Member, IEEE*, and Su Wang*, Student Member, IEEE, "Constrained Band Selection for Hyperspectral Imagery", IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, VOL. 44, NO. 6, JUNE 2006*

[4] C.-I. Chang and S. Wang, "Constrained band selection for hyperspectral imagery," *IEEE Transactions on Geosciences And Remote Sensing*, vol. 44, no. 6, pp. 1575-1585, 2006

[5] C.-I. Chang. *Hyperspectral Imaging: Techniques for Spectral Detection and Classification*. New York: Plenum, 2003

[6] I. U. Haq and X. Xu "A new approach to band clustering and selection for hyperspectral imagery," *IEEE ICSP Proceedings* 2008.

[7] M.E. Winter, "N-FINDR: an algorithm for fast autonomous spectral endmember determination in hyperspectral data," *Imaging Spectrometry V, Proc. SPIE 3753, pp. 266-277, 1999.*

[8] M.E. Winter, "A proof of the N-FINDR algorithm for the automated detection of endmembers in a hyperspectral image," *Algorithms and Technologies for Multispectral, Hyperspectral and Ultraspectral Imagery X, Proc. SPIE 5425, pp. 31-31, 2004.*

[9] D. Heinz and C.-I Chang, "Fully constrained least squares linear mixture analysis for material quantification in hyperspectral imagery," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 39, no. 3, pp. 529-545, March 2001.

[10] R. Huang and M. He, "Band selection based feature weighting for classification of hyperspectral data," *IEEE Geosci. Remote Sens. Lett.*, vol. 2, no. 2, pp. 156–159, Apr. 2005.